The stability of nonresponse rates according to socio-demographic categories

Zaletel, Metka; Vehovar, Vasja

Veröffentlichungsversion / Published Version
Konferenzbeitrag / conference paper

The Stability of Nonresponse Rates According to Socio-Demographic Categories

METKA ZALETEL AND VASJA VEHOVAR

Abstract: *The data from various sources (e.g. Central Population Register, Census data, Tax Register, Register of Territorial Units, etc.) were merged at the level of enumeration areas for the whole territory of Slovenia. The nonresponse and refusal rates of face-to-face surveys were connected to this model. The paper shows the stability of responses according to the given predictors.*

1 Introduction

During the last years, a lot of research has been conducted about reasons for the nonresponse in household surveys. Some of the results can be generalised to different countries, but some of them are really country - or region - specific. A typical example is the case of families with young children, which are generally more likely to respond; that was not proven to be true in Slovenia. On the other hand, there is a result which holds more generally: people living in the same neighbourhood or village do behave similarly when they get a request for the interview. Usually there is only one interviewer involved in a survey in a certain neighbourhood, so the interviewer's influence cannot be separated from the responding behaviour within a neighbourhood.

The paper analyses response rates at the level of primary sampling units and then constructs a model with explanatory variables at the same level, which enables us to predict response rates at the level of primary sampling units in the future. We first summarise the past research of nonresponse in Slovenia and its results in Section 2, then we explain the background and motivation of the present analysis in Section 3. In Sections 4 and 5 we first introduce all available data and then explain the methods used for the analysis; after this, the results are presented.

2 Past research and results

We have been searching for the reasons for nonresponse in Slovenia almost since the establishing of the country as an independent state in 1991. The nonresponse and refusal rates started to grow from the very beginning and, unfortunately, the growth has not stopped yet. In the case of the Labour Force Survey, we notice approximately 2% growth of nonresponse rate per year. With re-designing of the survey in 1997 and with employment of a small number of experienced interviewers we tried to stop the growth, but at the same time we introduced CAPI which - only in some regions - pushed it up again. Similar development is noticed in other surveys independent of the organisation - Statistical Office, academic organisations or marketing companies.

First, it was proven that substitutions cause biased estimations (Vehovar, 1993). At that time, substitutions were widely used in all kinds of surveys.

In 1995, the matching of persons selected in different samples with all available data sources was performed (Vehovar and Zaletel, 1995). The matching was possible because of the personal identification number that every person in Slovenia has. We matched the results of two surveys (Labour Force Survey and Household Budget Survey) with Census'91, Central Population Register, Database on Employed Persons, Taxation Register and Register of Unemployed Persons. We found that the results cannot be generalised across surveys since the Labour Force Survey gave us different predictors of response behaviour then the Household Budget Survey. Some interactions of two variables also appeared as predictors but they also cannot be generalised. The main predictors were education, income and type of dwelling. At the same time we estimated bias of some of the main estimates from both surveys. In some cases, the relative bias was up to 5% of the estimate.

In 1996, research of impact of confidentiality concern was made (Vehovar and Zaletel, 1996). The results are rather surprising: in general, people are not aware of confidentiality issues and possible abuse of survey data. The impact of confidentiality concern on nonresponse rates was thus not proven. There are certain topics which make some groups of people worried but we could not confirm that they are less likely to respond.

3 Motivation and background

Almost all samples of official surveys in Slovenia have the same sampling design - they are two stage stratified samples and primary sampling units are usually enumeration areas. The post-survey adjustment for nonresponse is also quite similar for most of the surveys: weights are calculated at the level of primary sampling units. If adjustment is done at the level of enumeration areas, perhaps we can also predict the nonresponse at the level of

enumeration areas. Another motivation for this idea are certainly the results from some of the countries (e.g. King, 1996) where the division of the country into small areas according to the socio-economic variables was made in advance. Then it was proven that the nonresponse rates vary across socio-demographic types of areas. We decided to generalise the idea: to build socio-demographic types of enumeration areas according to nonresponse rates achieved in some of the official surveys. This model would enable us to predict nonresponse rates for similar surveys in the future.

There are about 14,000 enumeration areas (EA) in Slovenia with 45 households each on average. Unfortunately, some of the EAs are very small or even empty, especially in remote areas. This fact caused a lot of problems in the process of sample designing and selection. In 1996, we merged all small EAs with their larger neighbours. We ended up with 9,872 clusters of enumeration areas (CEA) with an average of 65 households. The problem of small EAs vanished almost completely. Since 1996, primary sampling units in the majority of official surveys are CEAs.

4 Surveys and administrative data sources

To build a model for prediction of response rates at the level of enumeration areas, we selected some of the major (and largest) official surveys in Slovenia on one side to get data on achieved response rates at the level of enumeration areas and the register and census data on the other side to get socio-demographic data. In this section, we will first describe the surveys, then the administrative data and finally the variables selected from these rich sources of data.

4.1 Surveys

We included the following surveys:

- Labour Force Survey (LFS) 1994, 1995, 1996: this survey was conducted annually in May every year. Sample sizes were approximately 8,000 households per year. The whole fieldwork organisation was very similar from year to year: five follow-ups, advance letters, about 140 free-lance interviewers, face-to-face surveys in PAPI mode. Average length of interview was 18 minutes. The nonresponse rates were as follows: 8.9% in 1994, 9.0% in 1995 and 10.1% in 1996.
- Household Budget Survey (HBS) 1993, 1994, 1995, 1996: this survey was also conducted annually in December every year. The sample size in 1993 was 4,500 households, sample sizes in years from 1994 to 1996 were about 1,400 households. The fieldwork organisation was similar to that of the LFS with the exception of the

number of interviewers. In 1993, there were 109 free-lance interviewers. In later surveys, about 30 interviewers were involved. Average length of interview was about 90 minutes. The nonresponse rates were as follows: 19.7% in 1993, 17.8% in 1994, 18.0% in 1995 and 34.6% in 1996.

- Household Survey on Energy and Fuel Consumption (HSEFC): the survey was conducted for the first time in Slovenia in May 1997. The sample size was 5,000 households. The fieldwork was not organised by the Statistical Office of the Republic of Slovenia as for other surveys, but the organisation of the fieldwork was very similar. The number of interviewers was about 100. Average length of interview was 23 minutes. The nonresponse rate was 17.9%.

4.2 Administrative data sources

All major administrative data sources available at the Statistical Office of the Republic of Slovenia and some other organisations were used:

- Central Population Register (CRP)
- Census '91 database
- Database on Employed Persons in the Republic of Slovenia (DEP)
- Register of Territorial Units (RTU)
- Telephone Database (TD)

At this stage of research, the Taxation Register (maintained by the Ministry of Finance) has not been included in the estimations, but when the TR is available, the model will be re-estimated.

There are two important points which need to be stressed here. During the Census in 1991, there was a centroid determined for every building in Slovenia. Later, also the height above sea level of every building was estimated. According to these data, we defined the centroid for each CEA and for each settlement.

The second point concerns the time distance from the Census '91. All the data from the Census are obviously now 6 years old, but we took from the Census mostly data on dwellings and migrations. The Slovenian population is very stable and only about 2% of population is moving per year. In fact, most of those 2% are migrations within the same towns or villages. The situation concerning dwellings has not changed much in Slovenia since 1991 because not a lot of new dwellings have been build in-between. We can assume that Census data are good enough for our purposes.

4.3 Independent variables

First of all, we defined five sets of variables, concerning (1) persons, (2) dwellings, (3) households, (4) settlements and (5) clusters of enumeration areas. Then we re-calculated all these variables at the level of clusters of enumeration areas. We started the estimation of the model with the following variables:

Table 1: Available variables and their sources

		data source				
set	variable	CRP	Census	DEP	RTU	TD
1	proportion of children under 15 years					
1	proportion of persons over 65 years					
1	proportion of employed persons					
1	proportion of persons with higher education					
2	proportion of privately owned dwellings					
2	proportion of dwellings in apartment buildings					
2	proportion of weekend or summer houses					
3	proportion of farming households					
4	proportion of migration for school or work out of the settlement					
4	if the settlement is a centre of municipality or not					
4	type of settlement					
5	density of population					
5	air distance from centre of municipality					
5	telephone coverage					

4.4 Dependent variable

The first and natural idea for the selection of the dependent variable was the response rate at the level of enumeration areas. After merging all data we realised that there are some problems with data from Household Budget Survey 1993. We were able to define the initial sample size and responses for each CEA, but that was not the case for ineligible persons. In every survey we usually experience about 5% of ineligible households because of some differences between de-iure and de-facto addresses of those persons. After some investigation of the problem we concluded that this problem is equally spread all over the country and that results for the response rate and the completion rate (i.e.

number of responses divided by number of initial sample) are the same. So we simplified the problem and took the completion rate at the level of CEA as a dependent variable.

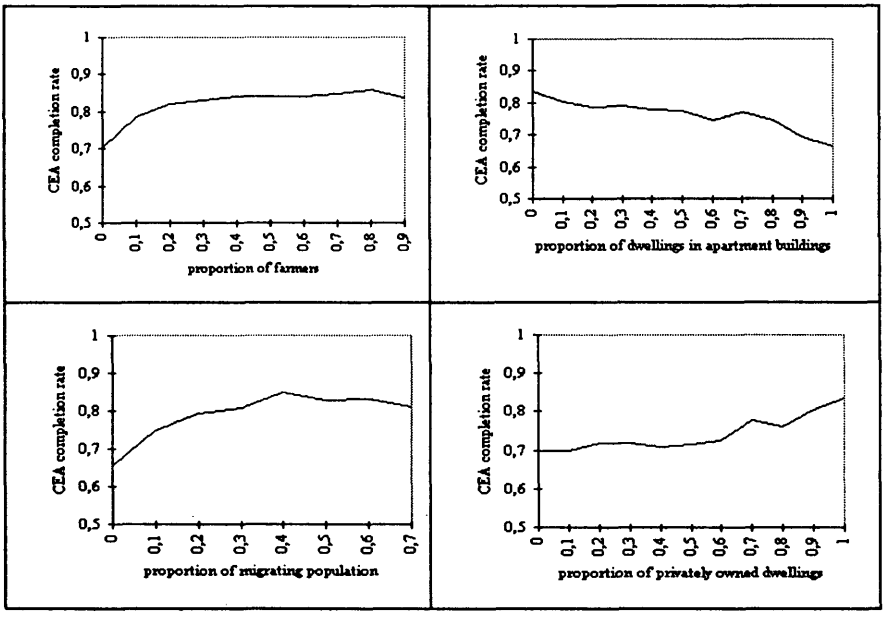
5 Analysis and results

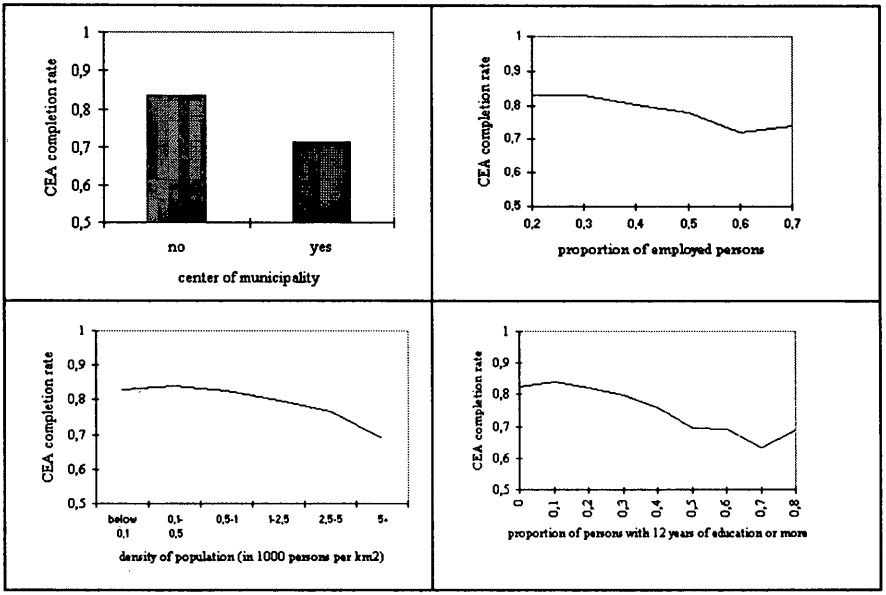
Firstly, we observed the dependence of completion rates on separate variables. Secondly, we took into account also the topic of the survey. Finally, the linear regression model was estimated.

5.1 Completion rates

Let us first observe a few figures presenting the dependence of CEA completion rates on selected variables. We calculated general completion rates irrespective of the survey.

Figure 1: CEA completion rates according to selected variables





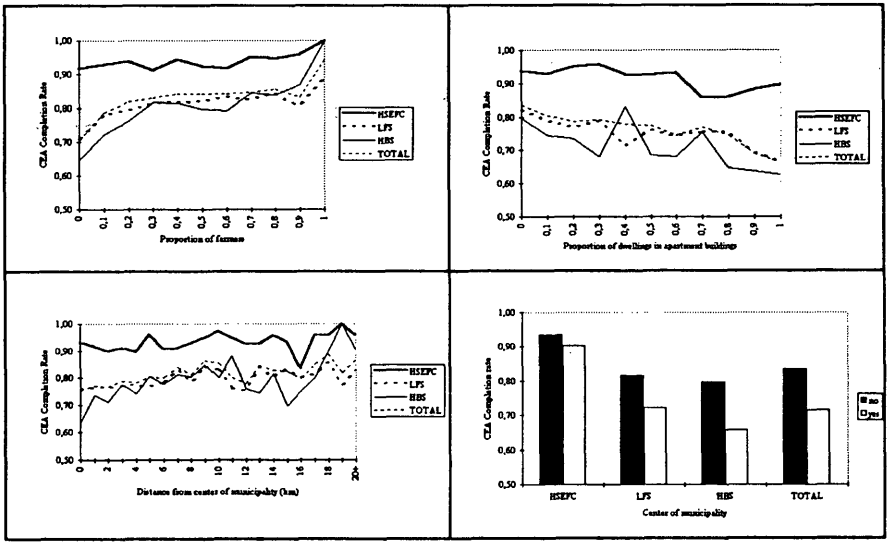
We can see that there exists a dependence of completion rates on most of the displayed variables. Here we do not present any numerical tables showing the dependence. There are two questions appearing right away:

1. Are the presented results survey dependent?
2. Is there any interaction between two variables?

5.2 Completion rates across different surveys

We notice that the HSEFC is behaving very differently in comparison with the other two surveys which are very similar. The same picture would be given with other variables which are not shown here. Even before the estimation of the model we can expect that we have to estimate separate models for each of the surveys included. At the same time we can say that the model for the HSEFC will not explain a lot of variability in completion rates. But let us first have a look at the estimation of the models.

Figure 2: CEA completion rates across different surveys



5.3 Regression model

The estimation of the regression model has shown what we expected and predicted according to the results of the previous section: the results cannot be generalised independently of the survey topic. Another result seen on the figures above was proved: available variables do not explain the variability in completion rates for the HSEFC at all.

In the table below we labelled the variables which were significant in the regression model. The level of significance is 0.05.

We can see that more or less the same variables are significant in the models for the LFS and the HBS. Only one variable is significant for the HSEFC, but even this one does not explain any variability of completion rates.

Table 2: Regression coefficients across different surveys

set	variable	HBS	LFS	HSEFC
1	proportion of children under 15 years			
1	proportion of persons over 65 years			
1	proportion of employed persons	-0.19		
1	proportion of persons with higher education	-0.20	-0.04	
2	proportion of privately owned dwellings			
2	proportion of dwellings in apartment buildings	-0.07	-0.11	
2	proportion of weekend or summer houses			
3	proportion of farming households		0.02	
4	proportion of migration for school or work out of the settlement	0.03		0.12
4	if the settlement is a centre of municipality or not	0.06	0.03	
4	type of settlement	-0.02	-0.02	
5	density of population			
5	air distance from centre of municipality			
5	telephone coverage			
	intercept	1.02	0.86	0.87

6. Conclusions and future work

The paper shows the modelling of completion rates at the level of clusters of enumeration areas according to given socio-demographic variables. We used data from three different official surveys and five different administrative sources. The achieved results somehow confirmed expectations we had according to similar research in the past:

- the results depend strongly on the survey topic and cannot be generalised;
- the key variables are education, type of dwelling and type of settlement;
- data on distances and population density are not important at this stage of research.

The research on enumeration areas will proceed with the taxation data added to the model. Income proved to be a very important explanatory variable in the past research. Data on the average number of contacts per household will also be added which will enable the estimation of costs of the survey in certain enumeration areas in advance.

References

Groves, R.M. (1989). Survey Errors and Survey Costs. New York. Wiley

Groves, R.M. and Couper, M.P. (1993). Correlates of Nonresponse in Personal Visit Surveys

King, J. (1996). The use of geo-demographic coding schemes for understanding household nonresponse. 7th International Workshop on Household Surveys Nonresponse. Rome, October 1996

Vehovar, V. (1993). Field Substitutions in Slovene Public Opinion Survey. Contributions to Methodology and Statistics. Ljubljana, FDV, pp. 39-66

Vehovar, V. and Zaletel, M. (1995). The Matching Project in Slovenia - Who are the Nonrespondents? Sixth International Workshop on Household Surveys Nonresponse. Helsinki, October 1995

Vehovar, V. and Zaletel, M. (1996). Does the confidentiality concern increase the nonresponse rate? 7th International Workshop on Household Surveys Nonresponse. Rome, October 1996